大规模知识处理与应用进展

王石 吴昱明 臧良俊 曹存根

摘要：本文分析了大规模知识处理的主要研究内容，包括知识表示、知识获取和知识应用的背景、研究现状和发展动态，概括介绍了中科院计算所在基于双层缺省逻辑的知识表示研究、面向自由文本的概念、关系、事件获取和验证研究，以及基于大规模知识库的企业智能客服系统的应用情况。

关键词：知识处理 本体学习 因果事件 常识知识 语义分析

1 引言

自 20 世纪 80 年代以来，人工智能领域的研究者经过了一系列的探索，逐渐认识到知识对于智能系统的重要性。莱纳特（Douglas B. Lenat）等提出的知识阈值理论[1]认为，一个系统之所以有智能是因为它具有可运用的知识，智能行为取决于知识的数量及其一般化的程度。目前，源于 Freebase 知识库的谷歌（Google）知识图谱已包含超过 2 亿用英文表述的实体和属性，成为支撑其众多应用的基础资源[2]。

中科院计算所智能信息处理重点实验室大规模知识处理课题组所建设的国家知识基础设施（Nation Knowledge Infrastructure，简称 NKI）[3]是一个庞大的、可共享的知识群体，它有三个目标：一是建立一个涉及多学科的专业知识库，目前包括医学[4]、军事[5]、地理[6]、植物[7]、数学[8]、考古[9]、民族、音乐、信息科学、农业、旅游、金融等学科的专业知识；二是建立领域无关知识以及常识知识库，目前已包括概念间上下位关系[10]、同指关系[11]、部分整体关系[12]，概念的属性及属性值[11]，还包括因果事件知识和常识性知识[13]；三是研究支持多种应用系统的知识应用方法。

在专业知识库和领域无关知识库的构建上，我们目前的主要研究方向集中于基于海量自由文本的本体学习和验证方面；在知识的应用方面，我们目前集中于自然语言的语义分析智能问答领域，成果已成功应用于建设多个行业的企业知识库和智能客服系统。本文，我们主要从这两个方面介绍国内外研究现状以及我们的进展。

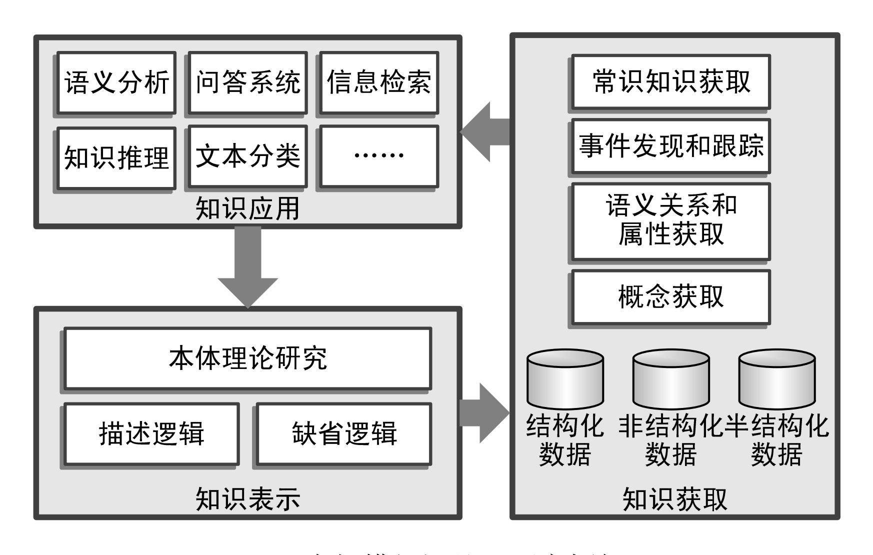


图1. 大规模知识处理研究框架

2　大规模知识处理国内外研究现状和中科院计算所研究进展

大规模知识处理的整体研究框架和我们的研究重点如图 1 所示。

在下面的三节中，我们将分别介绍知识表示、知识获取和知识应用的研究动态，以及我们的研究和应用进展。

2.1 基于双层缺省逻辑的知识表示研究

NKI（国家知识基础设施）是一个大规模的本体知识库，本体用于实现不同知识库之间的知识共享。NKI 采用多种形式语言表示，包括一阶逻辑、框架(frames)、描述逻辑(description logics)、RDF[1]/OWL[2]，以及缺省逻辑(default logic)。

框架是一种基于网络结构（network structures）的表示方法，用于表示个体的集合以及它们之间的关系[14]。然而，框架不具备形式语义，导致了不同框架系统具有不同的行为。为了克服这个缺陷，F-逻辑结合了基于框架的语言和基于逻辑的语言。

描述逻辑是一类（而非一种）知识表示语言。描述逻辑使用原子概念和原子角色通过概念构造子和角色构造子构建复杂概念和复杂角色。描述逻辑在表达能力和推理复杂度上有一个权衡（tradeoff）：简单的描述逻辑表达能力弱但推理复杂度低，复杂的描述逻辑表达能力强但推理复杂度高。描述逻辑在 OWL 中具有重要应用[15]。

缺省逻辑是实现非单调推理（non-monotonic or defeasible reasoning）的一种形式化方法。非单调性指已经得出的结论可能被后来的事实推翻，不能够保证单调性（即“知道的越多，得出的结论也就越多”）。例如，我们通常假定鸟是会飞的。当我们知道“Tweety[3]是一只幼鸟”时，就应该取消“Tweety 会飞”的结论。

我们在不同表示之间的翻译和缺省推理方面的研究包括：

1. 一阶理论、框架、概念图和描述逻辑具有相同的表达能力。为了证明两种表示具有相同的表达能力，需要在不同的表示之间进行转换，并且证明这种转换满足忠实性(faithful)和满足性(full)[16]；
2. 使用框架表示本体中的概念，用霍恩（Horn）逻辑程序作为自动推理；证明了对于本体和框架表示的知识库，基于霍恩逻辑程序的推理是正确的。
3. 描述逻辑中的概念之间具有包含关系$C \sqsubseteq D$，表示“如果x是概念C的实例，则x是概念D的实例”。除了包含关系，概念之间可能具有缺省包含关系$C \sqsubseteq_d D$，表示“如果x是概念C的实例，并且无法证明x不是概念D的实例，则x是概念D的实例”。为了自动推出概念之间的缺省包含关系，我们提出了双层缺省逻辑和双层缺省描述逻辑[17]。

2.2 面向自由文本的知识获取和验证

知识获取包括概念、关系、事件、常识知识的获取。就知识来源而言自由文本语料较之结构化和半结构化语料更易获得、规模更大，可以从中获得更丰富的知识。同时从中获得知识的任务也面临着更大的挑战。

2.2.1 概念获取和验证

1 Resource Description Framework，资源描述框架

2 Web Ontology Language，网络本体语言，一种用于描述语义网上本体论关系的语言，由国际组织 W3C 制定

3 一个卡通动画中的主角，有人译为“翠儿”或“崔弟”

概念是知识的基础元素之一，概念识别是文本知识获取的基础工作。中文概念识别和验证的困难在于：

1. 中文概念没有句法层上的普遍构造规律，存在构造歧义问题。例如“管理/v 思想/n”和“管理/v 公司/n”，两者具有相同的句法结构，但前者是指称概念而后者是述谓概念；
2. 不同于领域术语，概念本身没有明显的边界特征，在语料中也没有明显的邻接词特征，因此难以用基于边界的方法进行识别；
3. 汉语分词、未登录词识别、词性标注等基础性工作的困难和不足，也在一定程度上影响概念识别的效果。

相关研究方法主要分为基于语言学、基于统计和两者混合这样几个类别。基于语言学知识的概念识别方法在准确率上有非常明显的优点，但事先需要对语料库进行标注和分词，且分词和标注的效果对术语抽取结果有很大的影响[18]。这种方法可能对英语等西方语言更适合一些。与汉语相较，英语除了不存在分词问题外，还可以利用词态在词性标注上得到更好的效果。

基于统计方法的概念识别由于不考虑句法、语义上的信息，所以实现起来相对比较简单，并且这种方法不局限于某一专门领域，也不依赖任何外部资源。但是，这类工作不可避免地存在对低频术语的获取的困难和邻接高频词会引入噪声等问题，在准确率上往往难以保证[19][20]。规则和统计混合的方法主要结合上下文特征、句法结构规则以及统计信息来识别候选术语，这个方法越来越受到国内外研究者的重视[21]。

我们提出了一种基于概率句法构造模式、语义构造模式和邻接词统计特征三层递进的中文概念识别系统[22]。首先利用概念在句法层次上的构造规律进行识别，对于在句法上具有歧义的概念，进一步利用概念的语义构造规律进行处理。对于这两种方法都没有办法处理的概念，进一步根据概念的邻接词统计特征进行验证。其中，概念的语义构造特征利用了基于语料库和知识库的词汇间的语义相似度进行计算[23][24]。在基于关系抽取的任务测评中，这个方法可以达到约 93%的识别准确率，并且不受限于具体领域和概念类型。

2.2.2 关系和属性获取和验证

概念间的主要语义关系包括上下位、部分-整体、同指、地理位置等，每个概念也有其众多的属性及其对应的属性值。从自由文本中自动抽取和验证关系、属性、属性值是非常困难的。基于模式（Pattern）的语义关系获取方法近年来被广泛采用[25]，具有较高的精度。但这种方法离不开人工编辑的模式，并且获取的内容往往只局限于句子中的某个部分。引文[26]基于语料库，使用“核”（Kernel）的方法解决关系抽取问题，取得了较好的结果。其缺点是训练和预测的速度太慢，还不适于处理海量的数据。

我们已经进行了基于网络（Web）语料的上下位、部分-整体、同指关系获取。其中，在上下位关系获取方面，我们利用模式和概念空间中的关系验证方法，从大规模语料中抽取上下位关系[10]。我们首先给出一种利用词汇-句法模式获取候选上下位关系的方法，然后将候选上下位关系导入概念空间中，对概念空间中概念词和上下位关系进行分析，给出了上下位关系所具有的特征，并将这些特征以产生式规则的形式用于上下位关系的验证。据此，我们还设计了一种基于概念空间的启发式循环迭代验证模型。

在同指关系获取方面，我们提出一种多特征约束的方法，从语义受限的语料中获取同指关系[11]。方法分为候选同指关系的提取和验证两个阶段。首先基于目标词在语料中的分布

特征，提出分布距离、模式同质和并列分布三种策略混合的方法提取候选同指关系。然后基于同指标志词的语义特征和同指关系的组织结构特征，提出同指关系本体和同指关系图，再将两者结合提出一种联合策略验证候选的同指关系。

相对于预定义的关系种类，开放的属性、属性值获取更加困难。中文属性名称主要包括数量型、定性型、角色型三种类型。目前，获取属性名称所依据的语料数据，主要包括基于结构化数据源的提取，如 Web 查询日志[27]；基于半结构化的 Web 网页的提取，如从网页表格或表单中提取[28]，从维基百科文章（Wikipedia Articles）中提取[29]；以及基于多数据源的提取[30]。基于结构化和半结构化数据源的方法因其语料结构规整简短，具有一定的规律性，针对性强，主要采用弱文法和统计的方式进行提取，具有较高的准确率，但由于数据源的规模有限，因此召回率普遍不高。基于多数据源的方法主要是将结构化与非结构数据交叉迭代起来获取，首先从结构化数据中获取准确率较高的结果作为种子属性，然后使用种子属性从非结构化文本中迭代获取更多的属性。这种方法相比单一语料来源，综合考虑了准确率和召回率，但获取方法相对更加复杂，且结果属性的好坏和属性类型极大程度上依赖于种子。

表1 地域类概念的获取结果

概念	候选属性数	准确率					
		不验证	置信度 1	置信度 2	置信度 3	前向	后向
中国	13768	76.9%	88.0%	88.0%	94.0%	94.4%	94.8%
英国	6343	74.8%	86.5%	86.4%	92.4%	91.7%	92.5%
朝鲜	2549	71.5%	87.6%	87.6%	91.6%	91.0%	91.9%
北京	7661	76.3 %	90.1%	90.0%	94.6%	94.9%	95.2%
荆州	864	69.1%	88.4%	88.2%	90.1%	90.5%	90.3%

我们提出了一种基于前后缀迭代的属性名称获取方法，语料来源于非结构化数据源。方法的每一步迭代分为两个步骤：第一个步骤是从现有属性集合中选择合适的前后缀，构造词汇-句法模式，从 Web 网页中提取候选属性；第二个步骤是：采用基于相似性的验证模型对候选属性进行验证以扩充现有属性集合。实验结果表明，我们提出的基于前后缀迭代的获取方法得到的初始结果也具有较高的准确率，经过验证后，准确率又有了较大的提升。表 1 和表 2 是经过一轮迭代后的获取结果，其中验证后的准确率也就是前 80%结果的准确率。

表2 商业主体类概念的获取结果

概念	候选属性数	准确率					
		不验证	置信度 1	置信度 2	置信度 3	前向	后向
中石油	1854	73.7%	84.6%	84.7%	91.7%	92.0%	92.2%
摩根大通	440	66.4%	83.0%	82.7%	87.9%	85.3%	86.4%
碧桂园	837	69.4%	84.5%	84.5%	88.9%	88.8%	89.0%
比亚迪	1876	75.1%	88.1%	88.3%	92.0%	92.7%	93.1%
沃尔玛	2645	78.2%	86.0%	86.1%	92.6%	92.2%	92.8%

2.2.3 因果事件发现和推理

最初的因果知识获取研究主要借助基于知识的推理（Knowledge-based Inference）技术，并参照手工建立的领域知识库来完成。这些方法一般局限于特殊领域内的背景知识，甚至不能适用于除样本以外的其它数据。为克服早期工作的缺陷，其后的研究更多地借助语言特征来获取因果知识，主要包括基于统计和基于模式的方法，表 3、4 给出了因果关系获取的代表性工作结果。

表3 基于统计的因果关系获取的代表性相关工作

代表工作	统计特征	统计模型	知识源	结果
引文[31]	共现特征 论元共享特征	$P(v_1,v_2)$ $P_{arg}(n\|v_1)$, $P_{arg'}(n\|v_2)$	封闭语料	*P*: 0.45
引文[32]	共现特征 论元共享特征 单个动词分布特征	$P(v_1,v_2)$ $P_{arg}(n\|v_1)$, $P_{arg'}(n\|v_2)$ $P(v_1)$	封闭语料	*P*: 0.60
引文[33]	共现特征 时序特征	$PMI(v_1,v_2)$ $P(v_1 \rightarrow v_2)/P(v_2 \rightarrow v_1)$	电影剧本	*SRC*: 0.497
引文[34]	共现特征 时序特征 距离特征	$PMI(v_1,v_2)$ $P(v_1,v_2)$ $Distance(v_1,v_2)$	封闭语料	*F*: 0.60

表4 基于模式的因果关系获取的代表性相关工作

代表工作	模式	事件	分类器	知识源	结果
引文[35]	图模式	谓词+对象	——	英文领域文本	*P*:60%
引文[36]	$<NP_1\ Verb\ NP_2>$	名词短语	C4.5	英文文本	*P*:65%
引文[37]	因果动词	名词短语	C4.5	英文文本	——
引文[38]	因果连词	动词短语	SVM	日文文本	*P*: 85%
引文[39]	动词对	动词短语	SVM, NB	泰语文本	*P*: 88%; *R*:75%

我们提出了基于网络查询模式的因果关系抽取方法[13]。借助搜索引擎从网络中发现因果关系，将因果关系抽取任务划分为以下两个子任务：（1）因果关系的识别：基于表示因果关系的网络查询模式，从网络中抽取相关的文本。查询模式是因果知识在语言中表达方式的概括和抽象，包含原因事件槽和结果事件槽，以及它们的上下文约束；（2）因果关系的提取：从查询项匹配的文本中提取原因事件和结果事件，它们是对因果事件的自然语言描述。一般来说，封闭文本的模式匹配过程就是槽值的提取过程。但是在面向网络的关系提取中，查询项匹配是借助搜索引擎完成的，因此需要划分因果事件的边界。我们借助于依存句法分析，从语料中获取到事件五元组 E=(P, S, O, M, C)，即谓词（Predicate）、主体（Subject）、客体（Object）、述体（Modifier）、补体（Complement）五元组。

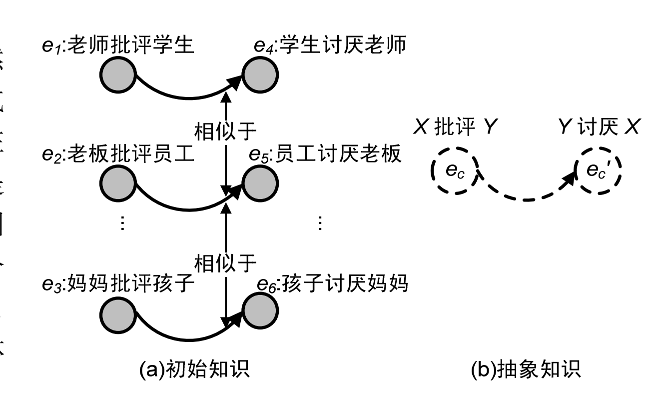


图2. 因果关系图的抽象和扩充

在生成因果关系图之后，我们又进行了以其为基础的知识扩充研究。理想的因果关系图是一个强连接的网络：事件之间关联紧密，因果关系能够连接成因果知识链。但实际获取到的因果关系一般相对独立。这是因为，许多事件的发生具有偶然性，其参与的因果关系较少；其次由于抽取模式的限制以及语料的稀疏问题，导致丢失了一些关系。因此，对于较孤立的

事件，我们希望进一步获取其相关的因果知识。我们采用了基于类比学习的方法，根据已有的因果关系扩充更多的关系，如图 2 所示。

2.3 大规模知识库在企业智能客服中的应用

在大规模知识处理应用方面，我们集中于基于大规模知识库的企业智能客服系统研发。近年来，我们已经将大规模知识库、知识管理技术、语义分析技术的研究成果应用于通信行业的移动、电信、联通等大运营商的数个省公司智能客服系统中，以及中国电信集团号百集团的综合信息服务、某大型 IT 企业的计算机硬件自助排障服务、金融领域知识培训和客服流大数据挖掘、面向特定领域的精准舆情监控等应用中，对大规模知识处理的研究方法进行了验证。

其中，应用最广泛的企业智能客服系统主要包括两大模块，一是基于本体理论的企业知识建模和管理，二是对用户咨询的自然语言语义分析。其中，语义分析是将自然语言分析成为可供计算机操作的逻辑表达式的过程，是自然语言处理的重要目标和难题，可以进一步直接应用于更多的应用系统，如图 3 所示。

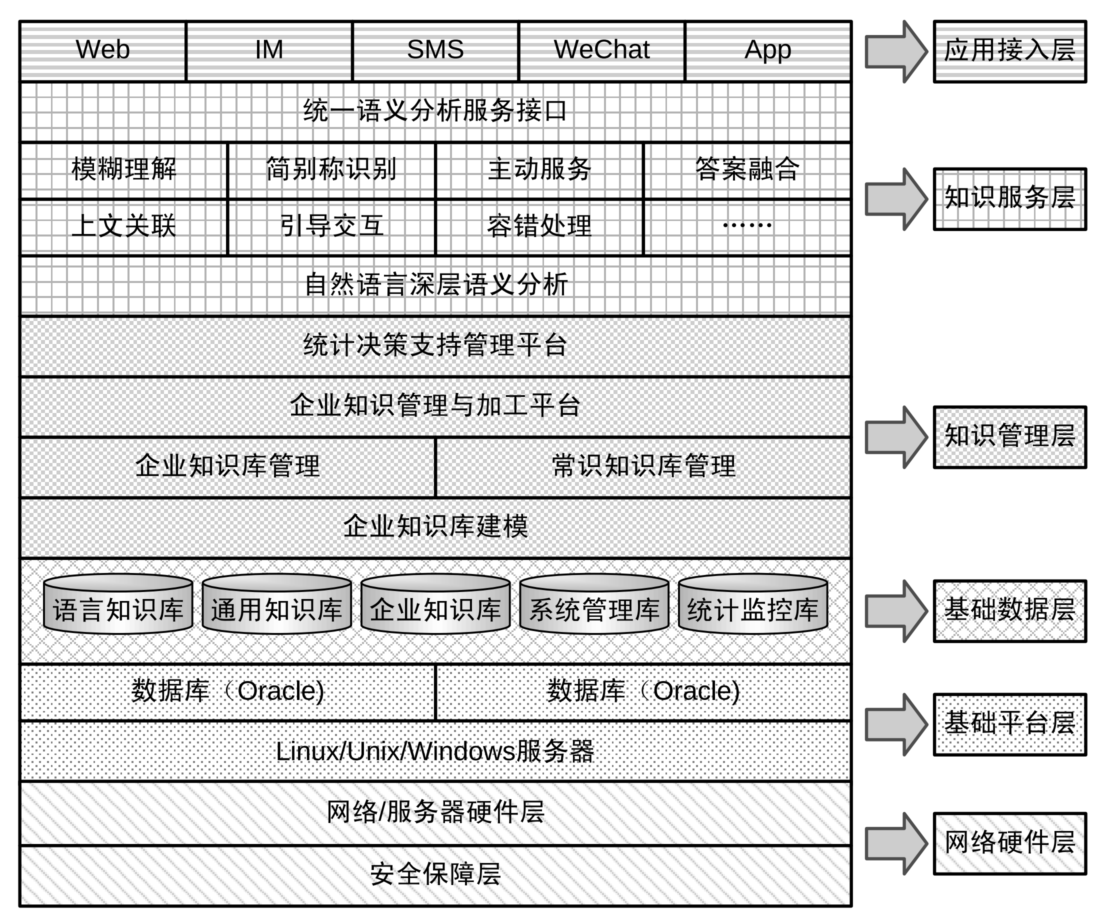


图3.　智能客服系统的体系结构

从图 3 中可以看出，大规模知识库是支撑整个系统的关键资源，对上层基于自然语言语义分析知识服务是不可缺少的。除此之外，我们考察了 Language Computer Corporation（语言计算机公司）、日本电信电话公司(NTT)、台湾中研院的 ASQA 系统等多个语义分析和智能问答系统，结合自身的研究和产品化过程，解决了自然语言理解的一些其他难题，包括缩略语识别技术、容忍形似/音似错别字的分词技术、用户意图模糊理解技术、用户咨询特征术语识别技术、简/别/俗称自动识别等，满足了实际应用的需要。

除了应用于智能客服之外，基于本体理论的企业知识管理也逐渐为企业决策支持、人员培训、产品运营、精确营销等提供越来越多的支撑。

3 总结和展望

近年来，我们在大规模知识表示、知识获取、知识应用方面进行了系统的研究和探索，取得了一定的进展。本文对相关工作做了一个简要概括性的说明。我们的下一步工作集中于两个方面：一是在大规模知识库的支撑下，进一步研究篇章、句子语义的语义文法学习，以此减少知识获取和智能问答应用中的人工干预工作量；二是研究"人们具备的有用的日常知识"，即常识性知识的表示、获取和应用。

参考文献：

[1] D. B. Lenat and E. A. Feigenbaum, "On the Thresholds of Knowledge," *IJCAI,* pp. 1173-1182, 1987.

[2] "Freebase data dumps," Google, 9 6 2013. Available: https://developers.google.com/freebase/data.

[3] C. Cao, Q. Feng and Y. Gao, "Progress in the development of national knowledge infrastructure," *Journal of Computer Science and Technology,* vol. 17, pp. 523 - 534, 2002.

[4] C. Cao, "Medical Knowledge Acquisition from the Electronic Encyclopedia of China," *Lecture Notes In Computer Science,* vol. 2101, pp. 268 - 271, 2001.

[5] Y. Lei and C. Cao, "Acquiring Military Knowledge from the Encyclopedia of China," *Proceedings of the Sixth International Conferences for Young Computer Scientists,* pp. 368-372, 2001.

[6] Y. Zhang and C. Cao, "Acquiring Knowledge of Chinese Mountains from the Encyclopedia of China," *Proceedings of the Sixth International Conferences for Young Computer Scientists,* 2001.

[7] F. Gu, C. Cao and Y. Sui, "A Domain-Specific Ontology of Botany," *Journal of Computer and Science Technology,* vol. 19, pp. 238-248, 2004.

[8] Q. Zeng, C. Cao and G. Tian, "Methods for Mathematical Concept Knowledge Representation, Acquisition and Management within NKIMath," *Second International Conference on Knowledge Economy and Development of Science and Technology,* pp. 27-35, 2004.

[9] C. Zhang, C. Cao ,F. Gu, "A Domain-Specific Formal Ontology of Archaeology and its Application in Knowledge Acquisition and Analysis," *Journal of Computer and Science Technology,* 2004.

[10] 刘磊，概念和上下位关系的获取理论和方法研究，北京：中国科学院研究生院博士学位论文，2007.

[11] 田国刚，受限中文语料的自监督文本知识获取研究，北京：中国科学院研究生院博士学位论文，2007.

[12] X. Cao and C. ,. S. Cao, "Extracting Part-Whole Relations from Unstructured Chinese Corpus," *FSKD,* pp. 175-179, 2009.

[13] 曹亚男，面向 Web 语料的因果知识获取研究，中国科学院研究生院博士学位论文, 2012.

[14] M. Kifer, G. Lausen， J. Wu, "Logical foundations of object-oriented and frame-based languages," *Journal of the ACM,* 1995.

[15] D. L. McGuinness， H. Van, "OWL web ontology language overview," *W3C recommendation,* 2004.

[16] 眭跃飞，高颖 ， 曹存根, "NKI 中的本体, 框架和逻辑理论," *Journal of Software,* 2005.

[17] 臧良俊，双层缺省逻辑和双层缺省描述逻辑研究,中国科学院研究生院博士学位论文, 2013.

[18] S.-H. WU , H. Wen-Lian, "SOAT: a semi-automatic domain ontology acquisition tool from Chinese corpus," *Proceedings of the 19th international conference on Computational linguistics*, 2002.

[19] C. Nigel and T. Koichi, "Comparison of character-level and part of speech features for name recognition in biomedical texts," *Journal of Biomedical Informatics,* vol. 37, pp. 423-435, 2004.

[20] 刘群, 张华平, 俞鸿魁 and 程学旗, "基于层叠隐马模型的汉语词法分析," *计算机研究与发展,* vol. 8, pp. 1421-1429, 2004.

[21] 余蕾, 从大规模中文语料中获取和验证概念的研究, 北京: 中国科学院计算技术研究所硕士学位论文, 2006.

[22] 王石, 中文实体名称的识别和语义分析方法研究, 北京: 中国科学院研究生院博士学位论文, 2009.

[23] S. Wangs ,C. Cungen, "Measuring Taxonomic Similarity between Words Using Restrictive Context Matrices," *Fifth International Conference on Fuzzy Systems and Knowledge Discovery,* 2008.

[24] S. WANG , C. CAO, "Mapping Synsets in WordNet to Chinese," *Conf. on Intelligent Text Processing and Computational Linguistics(CICLing),* 2012.

[25] P. Pantel, D. Ravichandran and E. Hovy, "Towards Terascale Knowledge Acquisition," *Proceedings of Conference on Computational Linguistics (COLING-04),* pp. 771-777, 2004.

[26] A. Culotta and J. Sorensen, "Dependency tree kernels for relation extraction," *Proceedings of ACL-2004,* 2004.

[27] P. Marius , V. D. Benjamin, "What You Seek isWhat You Get: Extraction of Class Attributes from Query Logs," *IJCAI,* 2007.

[28] Y. Naoki , T. Kentaro, "Open-Domain Attribute-Value Acquisition from Semi-Structured Texts," *Proceedings of the 6th International Semantic Web Conference,* 2007.

[29] C. B. Wladmir, S. M. Edleno , S. S. Altigran, "A Self-Supervised Approach for Extraction of Attribute-Value Pairs from Wikipedia Articles," *Int'l Symp on String Processing and Information Retrieval,* 2010.

[30] R. Joseph , P. Marius, "Low-Cost Supervision for Multiple-Source Attribute Extraction," *Proceedings of 10th International Conference on Intelligent Text Processing and Computational Linguistics,* 2009.

[31] K. Torisawa, "An Unsupervised Learning Method for Commonsensical Inference Rules on Events," *roceedings of the Second CoLogNet-EIsNET Symposium,* 2003.

[32] K.Torisawa, "Acquiring Inference Rules with Temporal Constraints by Using Japanese Coordinated Sentences and Noun-Verb Co-occurrences," *Proceedings of the HLT/NAACL,* 2006.

[33] B. Brandon , G. Roxana, "Using a Bigram Event Model to Predict Causal Potential," *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing),* 2009.

[34] R. Mehwish , G. Roxana, "Another Look at Causality: Discovering Scenario-Specific Contingency Relationships with No Supervision," *Proceedings of IEEE 4th International Conference on Semantic Computing(ICSC),* 2010.

[35] S. Christopher, S. C. Khoo , N. Yun, "Extracting causal knowledge from a medical database using graphical patterns," *Proceedings of The 38th. Annual Meeting of The Association for Computational Linguistics (ACL),* 2000.

[36] G. Roxana, "Automatic Detection of Causal Relations for Question Answering," *Proceedings of the ACL,* 2003.

[37] A. Terada, A Study of Text Mining Techniques using Natural Language Processing, Japan: Tokyo Institute of Technology, 2003.

[38] I. Takashi, I. Kentaro , M. Yuji, "Acquiring causal knowledge from text using the connective

marker tame," *ACM Transactions on Asian Language Information Processing*, 2005.

[39] P. Chaveevan , A. Kawtrakul, "Mining Causality from Texts for Question Answering System," *TRANSACTIONS on Information and Systems*, 2007.

作者简介：

王　石： 中科院计算所智能信息处理重点实验室、助理研究员，wangshi@ict.ac.cn
吴昱明： 中科院计算所智能信息处理重点实验室、助理研究员
臧良俊： 中科院软件所计算机科学重点实验室、博士后
曹存根： 中科院计算所智能信息处理重点实验室、研究员，cgcao@ict.ac.cn